On Maximal Repetitions in Words *

Roman Kolpakov[1] and Gregory Kucherov[2]

[1] French-Russian Institute for Informatics and Applied Mathematics, Moscow University, 119899 Moscow, Russia, e-mail: roman@vertex.inria.msu.ru
[2] LORIA/INRIA-Lorraine, 615, rue du Jardin Botanique, B.P. 101, 54602 Villers-lès-Nancy France, e-mail: kucherov@loria.fr

Abstract. A (fractional) repetition in a word w is a subword with the period of at most half of the subword length. We study maximal repetitions occurring in w, that is those for which any extended subword of w has a bigger period. The set of such repetitions represents in a compact way all repetitions in w.
We first study maximal repetitions in Fibonacci words – we count their exact number, and estimate the sum of their exponents. These quantities turn out to be linearly-bounded in the length of the word. We then prove that the maximal number of maximal repetitions in general words (on arbitrary alphabet) of length n is linearly-bounded in n, and we mention some applications and consequences of this result.

1 Introduction

Repetitions (called also periodicities) play a fundamental role in many topics of word combinatorics, formal language theory and applications. Several notions of repetition has been used in the literature. In its simplest form, a repetition is a word of the form uu, commonly called a *square*. A natural generalization is to consider, instead of squares, arbitrary powers, that is words of the form $u^n = \underbrace{uu\ldots u}_{n}$ for $n \geq 2$. We call such repetitions *integer repetitions* (or *integer powers*). If a word is not an integer repetition, it is called *primitive*. Integer repetitions can be further generalized to *fractional repetitions*, that is words of the form $w = u^n v$, where $n \geq 2$ and v is a proper prefix of u. u is called a *root* of w. If u is primitive, quantity $n + \frac{|v|}{|u|}$ is called the *exponent* of w, and $|u|$ is the *period* of w. Considering repetitions with fractional exponent may turn to be very useful and may provide a deeper insight of combinatorial properties of words [Dej72,Lot83,MP92,MRS95,CS96,JP99].

* The work has been done during the first author's visit of LORIA/INRIA-Lorraine supported by a grant from the French Ministry of Public Education and Research. The first author has been also in part supported by the Russian Foundation of Fundamental Research, under grant 96–01–01068, and by the Russian Federal Programme "Integration", under grant 473. The work has been done within a joint project of the French-Russian A.M.Liapunov Institut of Applied Mathematics and Informatics at Moscow University

Depending on the problem, the difference between the above three notions of repetition may not be relevant (for example if one wants to check whether a word is repetition-free) but, as will be seen below, may be important. Besides, if one wants to find (or to count) all repetitions in a word, it must be specified whether all *distinct* repetitions are looked for (that is, their position in the word is not relevant) or all the occurrences of (possibly syntactically equal) repetitions. In this paper we will be mainly concerned with the latter case, and we will sometimes say *positioned repetitions* to underline this meaning.

When one considers (integer or fractional) repetitions in a word, it is natural to consider "maximal" ones, that is those which cannot be further extended to the right/left to a bigger repetition with the same period. However, the definition of maximality differs depending on whether integer or fractional repetitions are considered. In case of integer repetitions, this amounts to those repetitions u^k, $k \geq 2$, which are not followed or preceeded by another occurrence of u. In case of fractional repetitions, a maximal repetition is a subword $u^n v$ (v a prefix of u, $n \geq 2$) which cannot be extended *by one letter* to the right or to the left without changing (increasing) the period. For example, the subword 10101 in the word $w = 1011010110110$ is a maximal fractional repetition (with period 2), while the subword 1010 is not. Another maximal fractional repetitions of w are prefix 10110101101 (period 5), suffix 10110110 (period 3), prefix 101101 (period 3), and the three occurrences of 11 (period 1).

In this paper we study maximal positioned fractional repetitions that, for the sake of shortness, we will call simply *maximal repetitions*.[1]. Maximal repetitions are important objects as they encode, in a most compact way, all repetitions in the word. For example, if we know all maximal repetitions in a word, we can easily obtain all squares in this word, with both primitive and non-primitive roots.

The question "How many repetitions can a word contain?" is interesting from both theoretical and applicative perspective. However, one must specify carefully which repetitions are counted.

A word of length n contains $O(n \log n)$ positioned primitively-rooted squares. This follows, in particular, from Lemma 10 of [CR95] which asserts that a word cannot contain in its prefixes more than $\log_\phi n$ primitive-rooted squares which immediately implies the $n \log_\phi n$ upper bound (ϕ is the golden ratio). On the other hand, in [Cro81] it was shown that Fibonacci words contain $\Omega(n \log n)$ positioned squares. Since all squares in Fibonacci words are primitively-rooted, this proves that $O(n \log n)$ is the asymptotically tight bound. A formula for the exact number of squares in Fibonacci words has been obtained in [FS99]. Note that in contrast, the number of distinct squares in Fibonacci words and in general, the maximal number of distinct squares in general words (over an arbitrary alphabet) is linear in the length [FS99,FS98].

The situation is different if only distinct squares are counted. In [FS99], it is shown that the k-th Fibonacci word f_k contains $2(|f_{k-2}|-1) = 2(2-\phi)|f_k| + o(1)$

[1] maximal repetitions have been called *runs* in [IMS97], *maximal periodicities* in [Mai89], and *m-repetitions* in [KK98]

distinct squares (ϕ is the golden ratio). The number of distinct squares in general words of length n is bounded by $2n$ (for an arbitrary alphabet), that was shown in [FS98] using a result from [CR95]. It is conjectured that this number is actually smaller than n, at least for the binary alphabet. Thus, in contrast to positioned squares, the maximal number of distinct squares is linear.

In [Cro81], Crochemore studies positioned primitively-rooted maximal integer powers, that is those subwords u^k, $k \geq 2$, which are not followed or preceeded by another occurrence of u. Similar to positioned squares, the maximal number of such repetitions is $\Theta(n \log n)$. The lower bound easily follows from the $\Omega(n \log n)$ bound for positioned squares in Fibonacci words, as Fibonacci words don't contain 4-powers, and an occurrence of a 3-power is an extension of two square occurrences. Therefore, the number of maximal integer powers in Fibonacci words is at least half the number of positioned squares, and is then $\Theta(n \log n)$.

What happens if we count the number of maximal repetitions instead of integer powers or just squares? Note that a word can contain much less maximal repetitions than maximal integer powers: e.g. if v is a square-free word over $\{a, b, c\}$, then word $v\#v\#v$ contains $|v| + 1$ (maximal) integer powers but only one maximal repetition. What is the maximal number of maximal repetitions in a word?

In the first part of the paper, we study maximal repetitions in Fibonacci words. The results of [IMS97] imply that Fibonacci words contain a linear number of maximal repetitions, with respect to the length of the word. This is showed, however, in an indirect way by presenting a linear-time algorithm which enumerates all maximal repetitions in a Fibonacci word. In this paper we first obtain directly the exact number of maximal repetitions in Fibonacci words, which is equal to $2|f_{k-2}| - 3$. Incidentally (or maybe not?), a Fibonacci word contains one less maximal repetitions than distinct squares.

We also estimate the sum of exponents of all maximal repetitions in a Fibonacci word. It is known ([MP92]) that Fibonacci words contain no subword of exponent greater than $2 + \phi$ but contain subwords of exponent greater than $2 + \phi - \varepsilon$ for every $\varepsilon > 0$. Therefore, from our previous result, the sum of exponents of all maximal repetitions is bounded from above by $(2+\phi)(2|f_{k-2}|-3) = 2(2-\phi)(2+\phi)|f_k| + o(1) = 2(3-\phi)|f_k| + o(1) \approx 2.764|f_k| + o(1)$. We could not obtain the exact formula for the sum of exponents, but we give a good estimation of it showing that this number is bounded asymptotically between $1.922 \cdot |f_k|$ and $1.926 \cdot |f_k|$.

Fibonacci words are known to contain "many" repetitions, and the fact that in Fibonacci words there is a linear number of maximal repetitions, rises the question if this is true for general words. We confirm this conjecture and prove that a word of length n over an arbitrary alphabet contains $O(n)$ maximal repetitions. The result is both of theoretical and practical interest. From the theoretical point of view, it contrasts to the above results about the $O(n \log n)$ number of positioned squares or integer repetitions, and shows that maximal repetitions are indeed a compact (linear) representation of all repetitions in a word. In par-

ticular, this answers the open question rised in [IMS97] whether all repetitions can be encoded in a linear-size structure and in particular, whether the number of maximal repetitions is linearly-bounded.

From the practical point of view, this result allows us to derive a linear-time algorithm of enumerating all maximal repetitions in a word. This algorithm, which is a modification of Main's algorithm [Mai89], will be briefly commented in the end of this paper, but will be presented in full details in an accompanying paper.

2 Definitions and Basic Results

Consider a word $w = a_1 \dots a_n$. Any word $a_i \dots a_j$ for $1 \le i \le j \le n$, which we denote $w[i..j]$, is a *subword* of w. A position in w is an integer number between 0 and n. Each position π in w defines a decomposition $w = w_1 w_2$ where $|w_1| = \pi$. The position of letter a_i in w is $(i-1)$. We say that subword $v = w[i..j]$ *crosses* a position π in w, if $i \le \pi < j$.

If w is a subword of u^n for some natural n, $|u|$ is called a *period* of w, and word u is a *root* of w. Clearly, p is a period of $w = a_1 \dots a_n$ iff $a_i = a_{i+p}$ whenever $1 \le i, i+p \le n$. Another equivalent definition is (see [Lot83]): p is a period of $w = a_1 \dots a_n$ iff $w[1..n-p] = w[p+1..n]$. The last definition shows that each word w has the minimal period that we will denote $p(w)$ and call often simply *the* period of w. The ratio $\frac{|w|}{p(w)}$ is called the *exponent* of w and denoted $e(w)$. Clearly, a root u of w such that $|u| = p(w)$, is *primitive*, that is u cannot be written as v^n for $n \ge 2$. Following [Lot83, Chapter 8], we call the roots u with $|u| = p(w)$ *cyclic roots*.

Consider $w = a_1 \dots a_n$. A *repetition* in w is any subword $r = w[i..j]$ with $e(r) \ge 2$. A *maximal repetition* in w is a repetition $r = w[i..j]$ such that

(i) if $i > 1$, then $p(w[i-1..j]) > p(w[i..j])$,
(ii) if $j < n$, then $p(w[i..j+1]) > p(w[i..j])$.

In other words, a maximal repetition is a repetition $r = w[i..j]$ such that no subword of w which contains r as a proper subword has the same minimal period as r. Note that any repetition in a word can be extended to a unique maximal repetition. For example, the repetition 1010 in word $w = 1011010110110$ extends to the maximal repetition 10101 obtained by one letter extension to the right.

A basic result about periods is the Fine and Wilf's theorem (see [Lot83]):

Theorem 1 (Fine and Wilf). *If w has periods p_1, p_2, and $|w| \ge p_1 + p_2 - gcd(p_1, p_2)$, then $gcd(p_1, p_2)$ is also a period of w.*

The following Lemma states some useful facts about maximal repetitions.

Lemma 2. *(i) Two distinct maximal repetitions with the same period p cannot have an overlap of length greater than or equal to p,*
(ii) Two maximal repetitions with minimal periods p_1, p_2, $p_1 \ne p_2$, cannot have an overlap of length greater than or equal to $(p_1 + p_2 - gcd(p_1, p_2)) \le 2\max\{p_1, p_2\}$.

Proof. Part (i) is easily proved by analyzing relative positions of two repetitions of period p and showing that if they intersect on at least p letters, at least one of them is not maximal. Part (ii) is a consequence of Fine and Wilf's theorem. If the intersection is at least $(p_1 + p_2 - gcd(p_1, p_2))$ long, then at least one of the cyclic roots of the two repetitions is not primitive, which is a contradiction.

3 Maximal Repetitions in Fibonacci Words

Fibonacci words are binary words defined recursively by $f_0 = 0$, $f_1 = 1$, $f_n = f_{n-1}f_{n-2}$ for $n \geq 2$. The length of f_n, denoted F_n, is the n-th Fibonacci number. Fibonacci words have numerous interesting combinatorial properties and often provide a good example to test conjectures and analyse algorithms on words (cf [IMS97]).

As it was noted in Introduction, Fibonacci word f_n contains $\Theta(F_n \log F_n)$ squares all of which are primitively-rooted. In [FS99], the exact number of squares in Fibonacci words has been obtained, which is asymptotically $\frac{2}{5}(3 - \phi)nF_n + O(F_n)$. Since general words of length n contain $O(n \log n)$ primitively-rooted squares [CR95], Fibonacci words contain asymptotically maximal number of primitively-rooted squares (at least up to a multiplicative constant).

In this section, we first count the exact number of maximal repetitions in Fibonacci words. Let R_n be the number of maximal repetitions in f_n. We prove the following

Theorem 3. *For all $n \geq 4$, $R_n = 2F_{n-2} - 3$.*

We follow the general proof scheme used in [FS99] for counting the number of positioned squares. Consider the decomposition $f_n = f_{n-1}f_{n-2}$ and call the position between f_{n-1} and f_{n-2} the *boundary*. Clearly, the maximal repetitions in f_n are divided into those which lie entirely in f_{n-1} or f_{n-2} and those which cross the boundary, that is intersect with f_{n-1} (call this intersection the left part) and with f_{n-2} (right part). We call the latter *crossing* repetitions. Note first that the left part and the right part of a crossing repetition cannot be both of exponent ≥ 2, since Fibonacci words don't have subwords of exponent 4. If either the left or the right part is of exponent ≥ 2, then the crossing repetition is an extension of a maximal repetition of respectively f_{n-1} or f_{n-2}. This implies that the only new crossing repetitions of f_n that should be counted are those that don't have their right and left part of exponent ≥ 2. Denote $c(n)$ the number of such crossing repetitions that we will call *composed* maximal repetitions of f_n. Then

$$R_n = R_{n-1} + R_{n-2} + c(n). \tag{1}$$

The following argument gives the solution.

Lemma 4. *For all $n \geq 8$, $c(n) = c(n-2)$.*

Consider the representation

$$f_n = f_{n-1}|f_{n-2} = f_{n-2}f_{n-3}|f_{n-3}f_{n-4} = f_{n-2}[f_{n-3}|f_{n-4}]f_{n-5}f_{n-4} \tag{2}$$

where | denotes the boundary, $n \geq 5$, and square brackets delimit the occurrence of f_{n-2} with the same boundary as for the whole word f_n. It is known that every repetition in Fibonacci words has the period F_k for some k (this is mentioned in [FS99] as a "folklore" result, proved in [Séé85]). Since $F_{n-3} > F_{n-4} > 2F_{n-6}$, it follows from (2) that if a composed maximal repetition of f_n has the period F_k for $k \leq n-6$, then it is also a composed maximal repetition of f_{n-2} and therefore is counted in $c(n-2)$. Vice versa, every composed maximal repetition of f_{n-2} with period F_k for $k \leq n-6$, is also a composed maximal repetition of f_n. We now examine the maximal repetitions of f_n with periods F_{n-2}, F_{n-3}, F_{n-4}, F_{n-5} which cross the boundary.

Crossing repetitions with period F_{n-2}. The last term of (2) shows that square $(f_{n-2})^2$ is a prefix of f_n that crosses the boundary. As $F_{n-1} < 2F_{n-2}$, the corresponding maximal repetition does not have a square in its left or right part and therefore is composed for f_n. Since $F_{n-2} > F_n/3$, any two maximal repetitions of f_n with period F_{n-2} intersect by more than F_{n-2} letters. By Lemma 2(i), this shows that f_n has only one maximal repetition with period F_{n-2}. Trivially, the maximal repetition under consideration is not a maximal repetition of f_{n-2}.

Crossing repetitions with period F_{n-3}. From the decomposition $f_n = f_{n-2}f_{n-3}|f_{n-3}f_{n-4}$ (see (2)), there is a square $(f_{n-3})^2$ with the root length F_{n-3} crossing the boundary. The corresponding maximal repetition does not extend to the left of the left occurrence of f_{n-3}, as the last letters of f_{n-3} and f_{n-2} are different (the last letters of f_i's alternate). Therefore, this maximal repetition does not have a square in its left or right part, and thus is composed for f_n. As this maximal repetition has a period both on the left and on the right of the boundary, it is the only maximal repetition with period F_{n-3} crossing the boundary (see Lemma 2(i)). Again, from length considerations, it is not an maximal repetition of f_{n-2}.

Crossing repetitions with period F_{n-4}. As $f_n = f_{n-2}[f_{n-3}|f_{n-4}]f_{n-5}f_{n-4} = f_{n-3}f_{n-4}[f_{n-4}f_{n-5}|f_{n-5}f_{n-6}]f_{n-5}f_{n-4} = f_{n-3}f_{n-4}[f_{n-4}\underbrace{f_{n-5}|f_{n-6}}_{f_{n-4}}f_{n-7}f_{n-6}]f_{n-5}f_{n-4}$ for $n \geq 7$, this reveals a maximal repetition of period F_{n-4} which crosses the boundary. However, this is not a composed maximal repetition of f_n, as it has a square on the left of the boundary. On the other hand, the restriction of this maximal repetition to f_{n-2} (subword in square brackets) *is* a composed maximal repetition for f_{n-2}.

It can be shown that this is the only maximal repetition of period F_{n-4} crossing the boundary. (There is another one which touches the boundary from the right, but does not extend to the left of it.) In conclusion, there is one composed maximal repetition of period F_{n-4} in f_{n-2} and no such maximal repetition in f_n.

Crossing repetitions with period F_{n-5}. Rewrite $f_n = f_{n-2}[f_{n-4}f_{n-5}|f_{n-5}f_{n-6}]f_{n-5}f_{n-4}$ which shows that there is a square of root length F_{n-5} crossing the boundary. Since the boundary is the center of this square, the latter corresponds to the only maximal repetition with period F_{n-5}

crossing the boundary. However, this maximal repetition is not a composed maximal repetition for f_n, as it has a square in its right part, as shown by the following transformation: $f_n = f_{n-2}[f_{n-4}f_{n-5}|f_{n-5}f_{n-6}]f_{n-6}f_{n-7}f_{n-4} = f_{n-2}[f_{n-4}f_{n-5}|f_{n-5}\underbrace{f_{n-6}]f_{n-7}}_{f_{n-5}}f_{n-8}f_{n-7}f_{n-4}$ for $n \geq 8$. On the other hand, the restriction of this maximal repetition to f_{n-2} (subword in square brackets) *is* a composed maximal repetition for f_{n-2}. Thus, there is one composed maximal repetition of the period F_{n-5} in f_{n-2} and no such maximal repetition in f_n.

In conclusion, two new composed maximal repetitions arise in f_n in comparison to f_{n-2}, but two composed maximal repetitions of f_{n-2} are no more composed in f_n, as they extend in f_n to form a square in its right or left part. This shows that $c(n) = c(n-2)$ for $n \geq 8$ and proves the Lemma.

A direct counting shows that $R_0 = 0,\ R_1 = 0,\ R_2 = 0,\ R_3 = 0,\ R_4 = 1,\ R_5 = 3,\ R_6 = 7,\ R_7 = 13$. Therefore, $c(3) = 0,\ c(4) = 1,\ c(5) = 2,\ c(6) = 3,\ c(7) = 3$. Since $c(n) = c(n-2)$ for all $n \geq 8$, then $c(n) = 3$ for all $n \geq 6$. We then have the recurrence relation $R_n = R_{n-1} + R_{n-2} + 3$ for $n \geq 6$ with boundary conditions $R_4 = 1,\ R_5 = 3$. Resolving it, we get $R_n = 2F_{n-2} - 3$ for $n \geq 4$. Theorem 3 is proved.

Thus, in contrast to squares, the number of maximal repetitions in Fibonacci words is linear. Using the same approach, we now estimate the sum of exponents of all maximal repetitions in f_n. A direct consequence of Theorem 3 and the fact that Fibonacci words don't contain exponents greater than $(2+\phi)$ [MP92], is that the sum of exponents is no greater, asymptotically, than $2(3-\phi)|f_k| \approx 2.764 \cdot |f_k|$. We now obtain a more precise estimation.

Denote $SR(n)$ the sum of exponents of all maximal repetitions in Fibonacci word f_n. We prove the following estimation for $SR(n)$.

Theorem 5. *$SR(n) = C \cdot |f_n| + o(1)$, where $1.922 \leq C \leq 1.926$.*

Similarly to (1), we write the recurrent relation

$$SR(n) = SR(n-1) + SR(n-2) + cx(n), \tag{3}$$

where $cx(n)$ is the sum of exponents of those left and right parts of crossing repetitions, which have the exponent smaller than 2. (If the exponent of the left or right part is 2 or more, it is counted in $SR(n-1)$ or $SR(n-2)$ respectively.) As before, the goal is to reduce $cx(n)$ to $cx(n-2)$, and a similar argument shows that for all crossing repetitions with the period F_k for $k \leq n-6$, nothing has to be done, as they occur completely inside f_{n-2} (see (2)) and are counted in $cx(n-2)$. As for Theorem 3, it remains to analyse repetitions with periods F_{n-2}, F_{n-3}, F_{n-4}, F_{n-5}.

The crossing repetition with period F_{n-2} is composed (both its left and right part is of exponent < 2), its length can be shown to be $F_n - 2 = F_{n-1} + F_{n-2} - 2$, and the exponent $\frac{F_{n-1}+F_{n-2}-2}{F_{n-2}}$. The crossing repetition with period F_{n-3} is also composed, of the length $2F_{n-3} + F_{n-4} = F_{n-2} + F_{n-3}$, and of the exponent

$\frac{F_{n-2}+F_{n-3}}{F_{n-3}}$. Let us turn to the crossing repetition with period F_{n-4}. Recall that it extends a repetition present in f_{n-2}. Its right part is of exponent < 2, and is inside f_{n-2}, therefore it is already counted in $cx(n-2)$, and it does not have to be added. Its left part is of exponent ≥ 2, and does not have to be counted in $cx(n)$. However, a part of it which is in f_{n-2} (namely $f_{n-4}f_{n-5}$), is of exponent < 2, and therefore has been counted in $cx(n-2)$. We then have to substract $\frac{F_{n-4}+F_{n-5}}{F_{n-4}} = \frac{F_{n-3}}{F_{n-4}}$. Similarly, the crossing repetition with the period F_{n-5} has the left part which is already counted in $cx(n-2)$, and the right part which should not be counted, but the part of it of exponent $\frac{F_{n-5}+F_{n-6}}{F_{n-5}} = \frac{F_{n-4}}{F_{n-5}}$ has been counted in $cx(n-2)$ and should be substracted. Putting everything together, we obtain the recurrence

$$cx(n) = cx(n-2) + 2 - 2/F_{n-2} + F_{n-1}/F_{n-2} + F_{n-2}/F_{n-3} - F_{n-3}/F_{n-4} - F_{n-4}/F_{n-5}, \quad (4)$$

for $n \geq 8$. Transforming further this expression, we obtain

$$cx(n) = n-1-2(1/F_{n-2}+1/F_{n-4}+...+1/F_4+1/F_2)+F_{n-1}/F_{n-2}+F_{n-2}/F_{n-3}$$

for even $n \geq 8$, and

$$cx(n) = n+1/2-2(1/F_{n-2}+1/F_{n-4}+...+1/F_3+1/F_1)+F_{n-1}/F_{n-2}+F_{n-2}/F_{n-3}$$

for odd $n \geq 9$. To join the cases, we rewrite (3) into

$$SR(n) = 2SR(n-2) + SR(n-3) + cx(n) + cx(n-1) = 2SR(n-2) + SR(n-3) + 2n - 3/2 - 2(\sum_{j=1}^{n-2} 1/F_j) + F_{n-1}/F_{n-2} + 2F_{n-2}/F_{n-3} + F_{n-3}/F_{n-4}.$$

The following estimation can be obtained using some elementary consideration.

$$-2(\sum_{j=1}^{n-1} 1/F_j) + F_n/F_{n-1} + 2F_{n-1}/F_{n-2} + F_{n-2}/F_{n-3} < 2,$$

for $n \geq 8$. We omit the proof. Using this estimation, we get that for all $n \geq 9$,

$$SR(n) \leq 2SR(n-2) + SR(n-3) + 2n + 1/2.$$

Solving this recurrence with initial conditions $SR(4) = 2, SR(5) = 6.5, SR(6) = 15\frac{11}{30}, SR(7) = 29\frac{27}{40}, SR(8) = 53\frac{142}{195}$, we obtain that

$$SR(n) \leq \frac{33}{520}(-1)^{n+1} + \frac{1}{\sqrt{5}}(40\frac{47}{130} - 25\frac{281}{1560}\bar{\phi})\phi^{n-6} + \frac{1}{\sqrt{5}}(40\frac{47}{130} - 25\frac{281}{1560}\phi)\bar{\phi}^{n-6} - n - \frac{15}{4} < \frac{1}{\sqrt{5}}(40\frac{47}{130} - 25\frac{281}{1560}\bar{\phi})\phi^{n-6} \approx 1.926 \cdot |f_n|.$$

The lower bound can be obtained as follows. A direct calculation gives the values $SR(23) = 1.922328 \cdot |f_{23}|, SR(24) = 1.922520 \cdot |f_{24}|$. Then using the obvious inequality $SR(n) \geq SR(n-1) + SR(n-2)$, we get $SR(n) \geq 1.922328 \cdot |f_n|$. Theorem 5 is proved.

4 Maximal number of maximal repetitions in a word

Since Fibonacci words contain "many" repetitions, Theorem 3 suggests the following question: Is it true that general words contain only a linear number of maximal repetitions? We answer this question affirmatively. We prove that the maximal number of maximal repetitions in words of length n is a linear function on n, regardless of the underlying alphabet. Denote by $Rep(n)$ the maximal number of maximal repetitions in words of length n (the alphabet is not fixed).

Theorem 6. $Rep(n) = O(n)$.

The proof of Theorem 6 is rather technical and cannot be given here because of space limitations. Actually, we prove that there exist absolute positive constants C_1, C_2 such that

$$Rep(n) \leq C_1 n - C_2 \sqrt{n} \log n$$

For the proof we refer the reader to [KK98].

5 Applications, Generalizations, Open Questions

In this concluding section we mention an important algorithmic application of Theorem 6, discuss its possible generalization, and formulate several related open questions.

An important application of Theorem 6 is that it allows to derive an algorithm which finds all maximal repetitions in a word in time linear in the length of the word.

The problem of searching for repetitions in a string (or testing if a string contains repetitions) has been studied since early 80's. Let us first survey known results. In early 80's, Slisenko [Sli83] claimed a linear (real-time) algorithm for finding all *distinct* maximal repetitions in a word. Independently, Crochemore [Cro83] described a simple and elegant linear algorithm for finding square in a word (and thus checking if a word is repetition-free). The algorithm was based on a special factorization of the word, called s-factorizarion (f-factorization in [CR94]). Another linear algorithm for checking whether a word contains a square was proposed in [ML85].

If one wants to explicitely list all squares (or integer powers) occurring in a word, there is no hope to do it in linear time, as their number may be of order $n \log n$. Several algorithms have been proposed in order to find all repetitions in time $O(n \log n)$. In 1981, Crochemore [Cro81] proposed an $O(n \log n)$ algorithm for finding all occurrences of primitively-rooted maximal integer powers in a word. Using a suffix tree technique, Apostolico and Preparata [AP83] described an $O(n \log n)$ algorithm for finding all positioned *right-maximal* fractional repetitions. Finally, Main and Lorentz [ML84] proposed another algorithm which actually finds all maximal repetitions in $O(n \log n)$ time. In 1989, using

Crochemore's s-factorization, Main [Mai89] proposed a *linear-time* algorithm which finds all *leftmost* occurrences of distinct maximal repetitions in a word.

As far as other related works are concerned, Kosaraju [Kos94] describes an $O(n)$ algorithm which, given a word, finds for each position the shortest square starting at this position. He also claims a generalization which finds all primitively-rooted squares in time $O(n + S)$ where S is the number of such squares. In [SG98a], Stoye and Gusfield proposed several algorithms that are based on a unified suffix tree framework. Their results are based on an algorithm which finds in time $O(n \log n)$ all "branching tandem repeats". In our terminology, branching tandem repeats are (not necessarily primitively-rooted) square suffixes of maximal repetitions. In a very recent paper, Stoye and Gusfield [SG98b] proposed a different approach, combining s-factorization (called Lempel-Ziv factorization in the paper) and suffix tree techniques. The goal achieved is to find, in linear time, a representative of each *distinct* square. The feasibility of this task is supported by the result of [FS98] mentioned in Introduction. The approach allows also to solve some other problems, e.g. to achieve the results claimed in [Kos94].

However, so far it has been an open question whether a *linear* algorithm for finding *all* maximal repetitions exists. In the concluding section of [Mai89], Main speculates that such an algorithm might exist. The same question is raised in [IMS97]. However, there has been no evidence in support of this conjecture as the number of maximal repetitions has not been known to be linear. Theorem 6 provides this argument. Using Theorem 6, it can be shown that Main's algorithm can be modified in order to find all maximal repetitions in linear time. This allows also to solve other related problems, e.g. to output all squares in a word in time $O(n + S)$, where S is the output size (cf [Kos94,SG98b]). The algorithm will be described in an accompaining paper. An interested reader may consult [KK98].

The results of this paper suggest an interesting question: Can Theorem 5 asserting the linearity of the sum of exponents of the maximal repetitions in Fibonacci words be also generalized to general words? Putting in direct terms, is the sum of exponents of maximal repetitions in a word also bounded linearly in the length of the word?

This conjecture is somewhat related to the hypothesis suggested in [SG98a] about the linearity of the maximal number of "branching tandem repeats" in a word. Branching tandem repeats are squares uu (not necessarily primitively-rooted) which are not followed by the first letter of u. To relate this to maximal repetitions, branching tandem repeats are suffixes of the maximal repetitions of length $2kp(r)$, where r is the corresponding maximal repetition and $k \geq 1$. The linearity of the maximal number of branching tandem repeats is stronger than our Theorem 6, as there are at least as many branching tandem repeats as maximal repetitions (each branching tandem repeat corresponds to a maximal repetition but one maximal repetition may contain several branching tandem repeats).

If the maximal sum of exponents of all maximal repetitions in a word were proved also linearly bounded, this would imply both our Theorem 6 and the

conjecture of [SG98a], and also shed some light on some facts we will mention below. Both authors of this paper strongly believe that this hypothesis is true. This is supported by computer experiments which show that in binary words that realize the maximal number of maximal repetitions, maximal repetitions are all of small exponent, typically not bigger than 3. This phenomenon is also illustrated by Fibonacci words, which contain "many" maximal repetitions, all of which are of exponent smaller than $2+\phi \approx 3.618$. The above hypothesis would shed light on this fact.

Let us make some other remarks about our results.

The main drawback of our proof of Theorem 6 is that it does not allow to extract a "reasonable" constant factor in the linear bound. It remains an open question if a simpler proof can be found which would imply a constant factor. We conjecture that for the binary alphabet this constant factor is equal to 1, which is supported by computer experiments.

Concerning counting results of Section 3, we note that Fibonacci words don't realize the maximal number of maximal repetitions among the binary words. For example, for length 21 this number is 15 (realized, e.g., by word 000101001011010010100) while Fibonacci word f_7 of length 21 contains 13 maximal repetitions.

While the number of maximal repetitions in Fibonacci words is one less than the number of distinct squares, computer experiments show that the maximal number of maximal repetitions in binary words of length n is apparently slightly bigger than the maximal number of distinct squares. In spite of this closeness between the number of maximal repetitions and that of distinct squares, there is no apparent connection between them. It is possible to conceive words with a big number of maximal repetitions and small number of distinct squares. For example, the result of [FS95] implies that there exist words with only three distinct squares but with unbounded number of maximal repetitions. Still, we are wondering if the fact that the number of maximal repetitions in Fibonacci words is one less than the number of distinct squares is a simple coincidence or it has some combinatorial explanation.

References

[AP83] A. Apostolico and F.P. Preparata. Optimal off-line detection of repetitions in a string. *Theoretical Computer Science*, 22(3):297–315, 1983.

[CR94] M. Crochemore and W. Rytter. *Text algorithms*. Oxford University Press, 1994.

[CR95] M. Crochemore and W. Rytter. Squares, cubes, and time-space efficient string searching. *Algorithmica*, 13:405–425, 1995.

[Cro81] M. Crochemore. An optimal algorithm for computing the repetitions in a word. *Information Processing Letters*, 12:244–250, 1981.

[Cro83] M. Crochemore. Recherche linéaire d'un carré dans un mot. *Comptes Rendus Acad. Sci. Paris Sér. I Math.*, 296:781–784, 1983.

[CS96] J.D. Currie and R.O. Shelton. Cantor sets and Dejean's conjecture. *Journal of Automata, Languages and Combinatorics*, 1(2):113–128, 1996.

[Dej72] F. Dejean. Sur un théorème de Thue. *J. Combinatorial Th. (A)*, 13:90–99, 1972.

[FS95] A.S. Fraenkel and J. Simpson. How many squares must a binary sequence contain? *Electronic Journal of Combinatorics*, 2(R2):9pp, 1995. `http://www.combinatorics.org/Journal/journalhome.html`.

[FS98] A.S. Fraenkel and J. Simpson. How many squares can a string contain? *J. Combinatorial Theory (Ser. A)*, 82:112–120, 1998.

[FS99] A.S. Fraenkel and J. Simpson. The exact number of squares in Fibonacci words. *Theoretical Computer Science*, 218(1):83–94, 1999.

[IMS97] C.S. Iliopoulos, D. Moore, and W.F. Smyth. A characterization of the squares in a Fibonacci string. *Theoretical Computer Science*, 172:281–291, 1997.

[JP99] J. Justin and G. Pirillo. Fractional powers in Sturmian words. Technical Report LIAFA 99/01, Laboratoire d'Informatique Algorithmique: Fondements et Applications (LIAFA), 1999.

[KK98] R. Kolpakov and G. Kucherov. Maximal repetitions in words or how to find all squares in linear time. Rapport Interne LORIA 98-R-227, Laboratoire Lorrain de Recherche en Informatique et ses Applications, 1998. available from URL: `http://www.loria.fr/~kucherov/res_activ.html`.

[Kos94] S. R. Kosaraju. Computation of squares in string. In M. Crochemore and D. Gusfield, editors, *Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, number 807 in Lecture Notes in Computer Science, pages 146–150. Springer Verlag, 1994.

[Lot83] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics and Its Applications*. Addison Wesley, 1983.

[Mai89] M. G. Main. Detecting leftmost maximal periodicities. *Discrete Applied Mathematics*, 25:145–153, 1989.

[ML84] M.G. Main and R.J. Lorentz. An $O(n \log n)$ algorithm for finding all repetitions in a string. *Journal of Algorithms*, 5(3):422–432, 1984.

[ML85] M.G. Main and R.J. Lorentz. Linear time recognition of square free strings. In A. Apostolico and Z. Galil, editors, *Combinatorial Algorithms on Words*, volume 12 of *NATO Advanced Science Institutes, Series F*, pages 272–278. Springer Verlag, 1985.

[MP92] F. Mignosi and G. Pirillo. Repetitions in the Fibonacci infinite word. *RAIRO Theoretical Informatics and Applications*, 26(3):199–204, 1992.

[MRS95] F. Mignosi, A. Restivo, and S. Salemi. A periodicity theorem on words and applications. In *Proceedings of the 20th International Symposium on Mathematical Foundations of Computer Science (MFCS)*, volume 969 of *Lecture Notes in Computer Science*, pages 337–348. Springer Verlag, 1995.

[Séé85] P. Séébold. Propriétés combinatoires des mots infinis engendrés par certains morphismes. Rapport 85-16, LITP, Paris, 1985.

[SG98a] J. Stoye and D. Gusfield. Simple and flexible detection of contiguous repeats using a suffix tree. In M. Farach-Colton, editor, *Proceedings of the 9th Annual Symposium on Combinatorial Pattern Matching*, number 1448 in Lecture Notes in Computer Science, pages 140–152. Springer Verlag, 1998.

[SG98b] J. Stoye and D. Gusfield. Linear time algorithms for finding and representing all the tandem repeats in a string. Technical Report CSE-98-4, Computer Science Department, University of California, Davis, 1998.

[Sli83] A.O. Slisenko. Detection of periodicities and string matching in real time. *Journal of Soviet Mathematics*, 22:1316–1386, 1983.